

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-078964

(43)Date of publication of application : 24.03.1998

(51)Int.Cl.

G06F 17/27  
G10L 3/00

(21)Application number : 09-162383

(71)Applicant : MICROSOFT CORP

(22)Date of filing : 19.06.1997

(72)Inventor : STEPHEN DAROU RICHARDSON  
GEORGE E HEIDOOON

(30)Priority

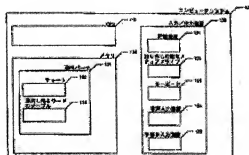
Priority number : 96 671203 Priority date : 25.06.1996 Priority country : US

## (54) METHOD AND SYSTEM FOR IDENTIFYING AND ANALYZING GENERALLY CONFUSED WORD BY NATURAL LANGUAGE PARSER

### (57)Abstract:

**PROBLEM TO BE SOLVED:** To perfectly parse a sentence including the words confused with each other by using a list of confused word sets and processing the word encountered in an input sentence to show a speech part that is shown in another word of the same set as the encountered word.

**SOLUTION:** An input/output device 120 of a computer system 100 includes a keyboard 123 and also optionally includes a voice input device 124 and a handwritten input device 125 which are used by a user to indirectly input a natural language text. A parser 131 of a memory 130 identifies and analyzes the words which are generally confused with each other. The parser 131 contains a chart 132 including a parse tree which shows the input text segments and the intermediate parsing results and also contains a table 133 of the generally confused words and the words that can be confused and mapped to the generally confused words (i.e., the intentional words).



特開平10-78964

(43) 公開日 平成10年(1998) 3月24日

| (51) Int.Cl. <sup>6</sup> | 識別記号  | 序内整理番号 | FI            | 技術表示箇所  |
|---------------------------|-------|--------|---------------|---------|
| G 0 6 F 17/27             |       |        | G 0 6 F 15/38 | D       |
| G 1 0 L 3/00              | 5 6 1 |        | G 1 0 L 3/00  | 5 6 1 G |
|                           |       |        | G 0 6 F 15/20 | 5 5 0 J |

審査請求 未請求 請求項の数10 OL (全 15 頁)

(21) 出願番号 特願平9-162383  
 (22) 出願日 平成9年(1997) 6月19日  
 (31) 優先権主張番号 08/671203  
 (32) 優先日 1996年6月25日  
 (33) 優先権主張国 米国 (US)

(71) 出願人 391055933  
 マイクロソフト コーポレイション  
 MICROSOFT CORPORATI  
 ON  
 アメリカ合衆国 ワシントン州 98052-  
 6399 レッドモンド ワン マイクロソフ  
 ト ウェイ (番地なし)  
 (72) 発明者 スティーヴン ダロウ リチャードソン  
 アメリカ合衆国 ワシントン州 98052  
 レッドモンド ノースイースト ワンハン  
 ドレッドアンドサードティセカンド 18028  
 (74) 代理人 弁理士 中村 稔 (外7名)

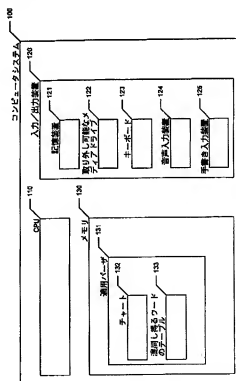
最終頁に続く

(54) 【発明の名称】 一般に混同するワードを自然言語パーザにおいて識別及び分析する方法及びシステム

## (57) 【要約】

【課題】 自然言語パーザにおいて一般に混同するワードを識別しそして分析する方法及びシステムを提供する。

【解決手段】 コンピュータシステムは、2つ以上のワードより成る入力テキストを、入力テキストのワードの中の1つのワードを含む潜在的に混同するワードからその意図されたワードへとマップする関係を使用してパズする。コンピュータシステムは、先ず、潜在的に混同するワードを含む入力テキストの各ワードに対して考えられるスピーチ部分を識別し、次いで、上記関係が潜在的に混同するワードをマップするところの意図されたワードに対して考えられるスピーチ部分を識別し、そしてそれらの識別されたスピーチ部分に構文的文法ルールを適用し、意図されたワードに対するスピーチ部分を含む完全な構文ツリーは発生されるが、潜在的に混同するワードに対するスピーチ部分を含む完全な構文ツリーは発生されないようにする。



## 【特許請求の範囲】

【請求項1】 コンピュータシステムにおいて1つ以上のワードを含む自然言語入力テキストのセグメントを文法ルールと複数のエントリーを含む辞書とを用いてパズする方法であって、各辞書エントリーは自然言語のワードに対応しそしてそのワードに対する1つ以上の考えられるスピーチ部分を指定し、上記方法は、

(a) 入力テキストセグメントを表すパズツリーと、それに対する中間のパズ結果とを含むチャートを作成し、

(b) 入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定されたスピーチの部分指定するスピーチ部分記録をそのワードに対して上記チャートに形成し、

(c) 入力テキストセグメントに生じるワードであって、別のワードと一般に混同するワードを識別し、

(d) その識別されたワードと一般に混同するワードに対する辞書エントリーで指定されたスピーチの部分指定するスピーチ部分記録を上記識別されたワードに対して上記チャートに形成し、そして

(e) 上記段階(b)及び(d)で形成された両方にスピーチ部分記録に文法ルールを適用する、という段階を備えたことを特徴とする方法。

【請求項2】 更に、一般に混同するワードのリストを使用し、このリストは、一般に混同するワードの各々に対し、そのワードと一般に混同するワードを含み、そして上記段階(c)は、入力テキストセグメントに生じるワードの1つを上記リストのワードの1つとマッチングさせる段階を含む請求項1に記載の方法。

【請求項3】 上記段階(b)は、入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定された考えられるスピーチ部分の各々を指定するスピーチ部分記録を上記チャートに形成し、そして上記方法は、更に、入力テキストセグメントの各ワードごとに、そのワードに対してチャートに形成されたスピーチ部分記録と一緒にリンクする段階を備え、1つ以上の文法ルールをスピーチ部分記録に適用することは、そのスピーチ部分記録がリンクされる他のスピーチ部分記録を検査することによりそのワードに対して他の考えられるスピーチ部分を決定することを含み、そして更に、上記方法は、段階(d)で形成されたスピーチ部分記録を段階(b)で識別されたワードに対して形成されたスピーチ部分記録にリンクさせる段階を含む請求項1に記載の方法。

【請求項4】 上記段階(d)は、上記段階(e)の実行を開始した後に行う請求項1に記載の方法。

【請求項5】 上記段階(d)は、上記段階(b)で形成されたスピーチ部分記録への文法ルールの適用が終了した後に行う請求項1に記載の方法。

【請求項6】 各文法ルール及びスピーチ部分記録は、

適用優先値が関連され、上記段階(e)は、適用優先値が減少する順に文法ルール及びリストのスピーチ部分記録を適用し、そして上記識別されたワードと一般に混同するワードに対するスピーチ部分記録に関連した適用優先値は、上記識別されたワードに対するスピーチ部分記録に関連した適用優先値より小さい請求項1に記載の方法。

【請求項7】 上記段階(e)の文法ルールの適用が、上記識別されたワードを含む入力テキストの完全なパズを形成しないが、一般に混同するワードを含む入力テキストの完全なパズを形成するときと、上記識別されたワードが上記一般に混同するワードと混同することを指示する段階を更に備えた請求項1に記載の方法。

【請求項8】 上記段階(e)の文法ルールの適用が、上記識別されたワードを含む入力テキストの完全なパズを形成しないが、又は一般に混同するワードを含む入力テキストの完全なパズを形成する場合には、自然言語センテンスが構文的に正しくないという指示を出力し、そして自然言語センテンスの識別されたワードが、その識別されたワードと一般に混同するワードと置き換えられた場合には、自然言語センテンスが構文的に正しいという指示を出力する、という段階を備えた請求項1に記載の方法。

【請求項9】 1つ以上のワードを含む自然言語入力テキストのセグメントを文法ルールと複数のエントリーを含む辞書とを使用してパズするための装置であって、各辞書エントリーは自然言語のワードに対応しそしてそのワードに対する1つ以上の考えられるスピーチ部分を指定し、上記装置は、

入力テキストセグメントを表すパズツリーと、それに対する中間のパズ結果とを含むデータ構造体と、入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定されたスピーチの部分指定するスピーチ部分記録を上記データ構造体に形成する一次スピーチ部分記録発生器と、

入力テキストセグメントに生じるワードであって、別のワードと一般に混同するワードを識別する識別子と、その識別されたワードと一般に混同するワードに対する辞書エントリーで指定されたスピーチの部分指定するスピーチ部分記録を、上記識別子で識別されたワードに対してチャートメモリに形成する二次スピーチ部分記録発生器と、

上記一次及び二次のスピーチ部分記録発生器で形成された両方のスピーチ部分記録に文法ルールを適用する文法ルール適用サブシステムと、を備えたことを特徴とする装置。

【請求項10】 ディスプレイ装置と、入力テキストセグメントが構文的に正しくないという指示を上記ディスプレイ装置に表示させると共に、入力テキストセグメントの識別されたワードが、その識別され

たワードと一般に混同するワードに置き換えられた場合には、入力テキストセグメントが構文的に正しいという指示を上記ディスプレイ装置に表示させるためのフィードバックシステムとを更に備えた請求項9に記載の装置。

#### 【発明の詳細な説明】

##### 【0001】

【発明の属する技術分野】本発明は、一般に、自然言語パーズングの分野に係り、より詳細には、自然言語テキストに生じるエラーを修正する技術に係る。

##### 【0002】

【従来の技術】文章を書くときに、筆者は、時々、あるワードが正しいところに誤って別のワードを使用することがある。例えば、筆者は、「ad」が正しいところにワード「add」を用いて次のような文章を書くことがある。The add convinced people. 互いに一貫して間違いを侵す「add」及び「ad」のようなワード対は、一般に混同すると言える。一般に混同するワードは、同様の発音を有する（例えば、「advise」対「advice」が、又は若干の文字の位置が異なる（例えば、「from」対「form」）ことが多い。上記例において、ワード「ad」は、「意図されたワード」即ち筆者により意図されたワードと称し、一方、ワード「add」は、「混同するワード」即ち筆者が意図されたワードと誤って置き換えたワードと称する。

【0003】混同するワードがセンテンスに含まれたときは、自然言語パーズは、センテンスをパーズングすることが困難である。自然言語パーズは、自然言語のセンテンスを分析して、センテンスの語彙及び構文内容を見分ける。例えば、チャートに基づく自然言語パーズは、入力センテンスの各ワードに対し辞書から辞書エントリを検索する。辞書エントリは、ワードに関する一般的情報を含む語彙記録と、ワードが表すスピーチの特定部分に特有の情報を含む参照用のスピーチ部分記録とを備えている。パーズは、1つ以上のスピーチ部分記録を、チャートと称する作用領域に入れ、それらはパーズングルールを受け、スピーチ部分記録が、より大きな構文単位、最終的にはセンテンスに合成される。

##### 【0004】

【発明が解決しようとする課題】自然言語パーズを使用し、意図されたワードとして筆者により意図されたスピーチの部分を持っていない混同するワードを含むセンテンスをパーズングするときには、自然言語パーズは、センテンスの完全なパーズを発生することができない。自然言語パーズの目的は、入力センテンスの意図された語彙及び構文内容を正確に表す完全なパーズを発生することであるから、混同するワードを含むセンテンスの完全なパーズを発生することのできる自然言語パーズが望まれる。

##### 【0005】

【課題を解決するための手段】本発明は、センテンスのような入力テキストセグメントのパーズング中に一般に混同するワードを識別しそして分析する自然言語パーズを提供する。一般に混同するワードを識別しそして分析する能力は、筆者により作成された文書に含まれたセンテンスの文法上の的確さを評価するためにワードプロセッサに関連して使用される文法チェッカーについて特に価値がある。しかしながら、当業者であれば、本発明は、自然言語パーズのいかなる用途にも有利に使用できることが明らかであろう。

【0006】本発明は、一般に混同するワードのセットのリストを使用する。各セットは、一般に混同する2つ以上のワードを含んでいる。本発明によるパーズは、入力センテンスにおいて遭遇したこれらセットの1つに現れるワードであって、あたかもそのセットの他のワードで表されるスピーチの部分を表し得るようなワードを処理する。例えば、ワード「add」及び「ad」が混同し得るワードのセットを構成しそしてワード「add」が入力センテンスに現れる場合に、パーズは、ワード「add」をあたかもそれが動詞又は名詞を表し得るかのようように処理する。というのは、ワード「add」は動詞を表し、そしてワード「ad」は名詞を表すからである。これは、遭遇するワードの辞書エントリにスピーチの対応部分がないか、又はスピーチの同じ部分が異なる数又は時制を有するようなセットの他のワードの辞書エントリからスピーチ部分記録をチャートに追加することを含む。これら付加的なスピーチ部分記録は、一般に、パーズが元のスピーチ部分記録にルールを適用する機会を有した後に、パーズングプロセスに後で追加される。

【0007】本発明の実施形態は、更に、一般に混同するワードを識別及び分析した結果を表示するために文法チェッカーのユーザインターフェイスを提供する。又、ある実施形態において、本発明は、これらの追加されたスピーチ部分記録へのレファレンスを、遭遇したワードの語彙記録に追加し、ワードの語彙記録を用いてワードがスピーチのどの部分を表すかを決定するためのルールが、遭遇したワードがスピーチのこれら追加部分を表す確率を考慮するようにする。

##### 【0008】

【発明の実施の形態】自然言語パーズにおいて一般に混同するワードを識別しそして分析する方法及びシステムが提供される。好ましい実施形態において、本発明は、一般に混同するワードのセットのリストを使用し、これは、ユーザにより修正することができる。本発明によれば、パーズは、入力センテンスにおいて遭遇したこれらセットの1つに現れるワードであって、あたかもそのセットの他のワードで表されるスピーチの部分を表し得るようなワードを処理する。例えば、ワード「add」及び「ad」が混同し得るワードのセットを構成しそしてワ

10

20

30

40

50

ード「add」が入力センテンスに現れる場合に、パーザは、ワード「add」をあたかもそれが動詞又は名詞を表し得るかの如く処理する。というのは、ワード「add」は動詞を表し、そしてワード「ad」は名詞を表すからである。これは、遭遇するワードにスピーチの対応部分がないセットの他のワードの辞書エントリからスピーチ部分記録をチャートに追加することを含む。これらの付加的なスピーチ部分記録は、一般に、パーザが元のスピーチ部分記録にルールを適用する機会を有した後に、パーズングプロセスに後で追加される。又、ある実施形態では、本発明は、これらの追加されたスピーチ部分記録へのレファレンスを、遭遇したワードの語彙記録に追加し、従って、ワードの語彙記録を用いてワードがスピーチの他のどの部分を表すかを決定するためのルールが、遭遇したワードがスピーチのこれら追加部分を表す確率を考慮するようにする。

【0009】図1は、パーザが好ましい動作する汎用コンピュータシステムの高レベルブロック図である。コンピュータシステム100は、中央処理ユニット(CPU)110と、入力/出力装置120と、コンピュータメモリ(メモリ)130とを備えている。入力/出力装置の中には、ハードディスクドライブのような記憶装置121がある。又、入力/出力装置は、取り外し可能なメディアドライブ122を含み、このドライブは、適用パーザを含むソフトウェア製品をインストールするのに使用でき、これらソフトウェア製品は、CD-ROMのようなコンピュータ読み取り可能な媒体に設けられる。更に、入力/出力装置は、ユーザが自然言語テキストを直接的に入力するのに使用するキーボード123も含む。又、入力/出力装置は、ユーザが自然言語テキストを間接的に入力するのに使用する音声入力装置124及び手書き入力装置125も任意に含む。音声入力装置を用いてユーザにより入力される自然言語テキストは、好ましくは、音声認識装置(図示せず)により音声データから変換される。同様に、手書き入力装置を用いてユーザにより入力される自然言語テキストは、好ましくは、手書き認識装置(図示せず)を用いて手書きデータから変換される。メモリ130は、一般的に混同するワードを識別及び分析するためのパーザ131を含む。パーザは、入力テキストセグメント及び中間パーズ結果を表すパーズツリーを含むためのチャート132を備えている。又、パーザは、一般に混同するワードを、それらと一般に混同するワード(即ち、意図されたワード)へとマップする混同し得るワードのテーブル133も備えている。例えば、ワード「add」は、ワード「ad」へとマップされ、筆者が「ad」ではなくワード「add」を間違えて使用するかもしれないことを指示する。混同し得るワードのテーブルは、記憶装置に記憶されてもよいし、又は取り外し可能なメディアドライブを用いて取り外し可能な媒体に記憶されてもよい。パーザは、

上記のように構成されたコンピュータシステムにおいて実施されるのが好ましいが、異なる構成のコンピュータシステムでも実施できることが当業者に明らかであろう。

【0010】適用パーザを用いて入力テキストをパーズしながら、一般に混同するワードを識別及び分析する一例を、図2ないし7に関連して説明する。図2は、入力テキストに生じる潜在的に混同するワードを含む入力テキストのワードに対しパーザがスピーチ部分記録(part-of-speech record)をチャートに追加するところを示すチャート図である。図2は、例示の入力ストリング201「The add convinced people.」を示している。図2は、更に、パーザが入力テキストに現れるワードに対するスピーチ部分記録をチャートに追加した後のパーザのチャート200の内容も示している。このチャートは、スピーチ部分記録211-215を含み、その1つ以上は、入力テキストに生じる各ワードを示す。スピーチ部分記録211は、ワード「the」を表し、スピーチ部分記録212は、ワード「add」を表し、スピーチ部分記録213は、ワード「convinced」を表し、そしてスピーチ部分記録214及び215は、ワード「people.」を表す。各スピーチ部分記録は、そのワードが表し得るスピーチの1つの考えられる部分の指示と、動詞のスピーチ部分記録に対する動詞の時制のような付加的な関連語彙情報とを含む。入力ストリングに生じるワードに対し辞書入力から検索されるのが好ましい。

【0011】図3は、チャートの内容により暗示されるルールをパーザが適用するところを示す。即ち、パーザは、チャートに既に存在する形式の記録を結合することのできるルールを適用する。図3は、動詞及び名詞を動詞句即ち「VP」に変換するルールの適用を示す。このルールの適用は、ワード「convinced」に対する動詞のスピーチ部分記録313を「people」に対する名詞のスピーチ部分記録314と結合する動詞句記録321を形成する。スピーチ部分記録及びルールにより形成された記録の両方は、付加的なルールを包含する。これら付加的なルールは、繰り返し適用される。ここに示す例の場合には、チャートの記録を結合するのに他のルールは首尾良く適用されない。

【0012】これらルールの適用により完全なパーズが発生された場合には(即ち、入力ストリングの全てのワードをカバーするセンテンス記録が形成された場合には)、パーズングが終了しそして完全なパーズが返送されるが、さもなくば、パーザの動作が続けられる。入力テキストが潜在的に混同するワードを含む場合には、パーザの動作が続けられるが、さもなくば、パーザは完全なパーズを発生することができず、欠陥を返送する。図4は、入力テキストにおいて潜在的に混同するワードを

識別するのに使用される例示的な混同し得るワードのテーブルを示す図である。この混同し得るワードのテーブル400は、潜在的に混同するワードの欄と、考えられる意図されたワードの欄を含む。各行において、潜在的に混同するワードの欄は、1つ以上の他のワードに対して混同することのあるワードを含む。その行において、考えられる意図されたワードの欄は、潜在的に混同するワードと混同し得る1つ以上の考えられる意図されたワードのリストを含む。例えば、行402は、ワード「add」がワード「ad」と混同し得ることを示す。行401は、その逆もあることを示し、即ちワード「ad」がワード「add」と混同することを示している。ワード間の幾つかの潜在的な混同は、方向性であり、即ちある対の一方のワードは、その対の他方のワードと潜在的に混同し得る(例えば、行403は、「can't」が「can't」と混同し得ることを示す)が、その逆はない(例えば、「can't」は、潜在的に混同するワードの欄に現れない)。行409ないし411は、潜在的に混同するワードが2つ以上の考えられる意図されたワードと混同し得ることを示す。パーザは、入力テキストのワードを、混同し得るワードのテーブルの潜在的に混同するワード欄のワードと比較する。入力ストリングのいずれかのワードが、潜在的に混同するワード欄のワードと一致する場合には、その入力テキストが潜在的に混同するワードを含む。

【0013】図5は、入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対しパーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。パーザは、潜在的に混同するワード以外のスピーチ部分を有する考えられる意図されたワードに対するスピーチ部分記録をチャートに追加するのが好ましい。例えば、パーザは、好ましくは、考えられる意図されたワード「ad」に対し名詞のスピーチ部分記録を追加する。というのは、そのスピーチ部分が、その潜在的に混同するワード「ad」に対して考えられるスピーチ部分とは異なるからである。又、パーザは、好ましくは、潜在的に混同するワードとは異なる時制を有する考えられる意図されたワードに対しスピーチ部分記録をチャートに追加する。例えば、パーザは、好ましくは、考えられる意図されたワード「lapse」に対し複数名詞のスピーチ部分記録を追加する。というのは、その数が、潜在的に混同するワード「lapse」の単数名詞と異なるからで

10

20

30

40

50

ある。更に、混同するワードの特定セットに対し、ユーザは、たとえスピーチ部分、時刻及び数が同じであっても、セット内のワードに対しスピーチ部分記録をチャートに追加することを指定するのが好ましい。図5は、パーザがワード「ad」に対する名詞のスピーチ部分記録516をチャートに追加したところを示す。というのは、行402に示すように、入力ストリングに現れるワード「ad」が、辞書が考えられるスピーチ部分として名詞を指定するところのワード「ad」と混同し得るからである。

【0014】図6は、考えられる意図されたワードに対してスピーチ部分記録をチャートに追加した後にチャートの内容により暗示されるルールをパーザが適用するところを示すチャート図である。図6は、チャートに記録622及び623を形成するルールをパーザが適用するところを示す。記録622は、「the」に対する冠詞のスピーチ部分記録611を、考えられる意図されたワード「ad」に対する名詞のスピーチ部分記録616と結合して、名詞句(NP)を形成する。記録623は、名詞句の記録622と動詞句の記録621を結合してセンテンスを形成する。又、記録623は、入力テキストの各ワードを表すリーフを有するツリーのヘッドノードを構成するという点で、入力ストリングの各ワードを「カバー」する。

【0015】別のワードに対するスピーチ部分記録を追加した後のチャートの内容によって暗示されたルールの適用が完全なパーズを形成した場合には、パーザは、的確な成功を返送するが、さもなくば、パーザは失敗を返送する。図6から明らかなように、この例では、パーザは、センテンス記録623が入力テキストの全てのワードをカバーし、それ故、的確な成功を返送するという点で、完全なパーズを形成している。

【0016】図7は、パーザを用いた文法チェッカーの視覚的ユーザインターフェイスを示すスクリーン図である。この文法チェッカーのユーザインターフェイスは、好ましくは、ウィンドウ700を表示する。ウィンドウ700は、好ましくは、現在チェックされているセンテンス710を含む。更に、ウィンドウは、現在センテンスの特定のワードがおそらく別のワードと混同しているという指示720を含む。又、ウィンドウは、好ましくは、潜在的に混同するワードを置き換えるという提案710も含む。更に、ウィンドウは、好ましくは、提案を受け入れて潜在的に混同するワードを置き換えることをユーザが選択できるボタン740と、提案を拒否しそして潜在的に混同するワードの置き換えを排除することをユーザが選択できるボタン750とを含む。

【0017】図8は、一般に混同するワードを識別しそして分析しながら入力テキストをパーズするために適用パーザにより好ましく実行される高レベルステップを示すフローチャートである。ステップ801において、パ

一々は、入力テキストに生じる潜在的に混同するワードを含む入力テキストのワードに対しスピーチ部分記録をチャートに追加する。ステップ802において、パーザは、チャートの内容により暗示されたルールの一つを適用する。ステップ803において、ステップ802のルールの適用により完全なパースが形成された場合には（即ち、入力ストリングの全てのワードをカバーするセンテンス記録が形成された場合には）、これらのステップは終了し、完全なパースが返送されるが、さもなくば、パーザはステップ804に続く。ステップ804において、パーシングが終了した場合、即ち暗示された全てのルールが適用されるか又は適用されたルールの全数が上限を越えた場合には、パーザは、ステップ805に続き、さもなくば、パーザは、ステップ802に続いて、暗示された別のルールを適用する。ステップ805において、入力テキストが潜在的に混同するワードを含む場合には、パーザは、ステップ806に続き、さもなくば、パーザは、完全なパースを形成することができず、失敗を返送する。入力テキストが潜在的に混同するワードを含むかどうか決定するために、パーザは、入力テキストのワードを、混同し得るワードのテーブルの潜在的に混同するワード欄のワードと比較する。入力ストリングのいずれかのワードが、潜在的に混同するワード欄のワードと一致する場合には、入力テキストが潜在的に混同するワードを含む。ステップ806において、パーザは、入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対し、スピーチ部分記録をチャートに追加する。ステップ807において、パーザは、別のワードに対しスピーチ部分記録をステップ806で追加した後にチャートの内容により暗示されたルールの一つを適用する。ステップ808において、完全なパースが発生された場合には、パーザは、的確な成功を返送し、さもなくば、パーザは、ステップ809に続く。ステップ809において、パーシングが完了した場合、即ち暗示された全てのルールが適用されるか又は適用されたルールの全数が上限を越えた場合には、パーザは、失敗を返送し、さもなくば、パーザはステップ807に続き、別の暗示されたルールを適用する。

【0018】幾つかの文法ルールは、各ワードに対してスピーチの特定部分に適用される間に、各ワードが表すスピーチの全ての考えられる部分を考慮する。このような文法ルールは、ルールを適用するのに必要であるが完全なパースの発生に貢献しそうな処理リソースの量を減少することができる。このようなリソースをサポートするために、スピーチ部分記録と一緒にリンクし、ワードに対して考えられる全てのスピーチ部分を容易に決定することができる。本発明の好ましい実施形態によれば、考えられる意図されたワードに対するスピーチ部分記録は、パーシングプロセスの始めに、潜在的に混同す

るワードに対するスピーチ部分記録にリンクされるのが好ましい。図8ないし11は本発明のこの特徴を示す。図9は、図2の別の図であって、ステップ801の実行の後であって且つルールを適用する前のチャートの内容を示す。入力テキストのワードにより表されたスピーチの潜在的な部分に関するデータを含むのではなく、スピーチ部分記録911ないし915は、この情報を含むデータ構造体へのポインタを含む。例えば、スピーチ部分記録914は、潜在的なスピーチ部分の名詞及び他の関連する語彙情報を含むスピーチ部分データ構造体971へのポインタを含む。別の好ましい実施形態（図示せず）によれば、スピーチ部分データ構造体へのポインタをデレファレンスする時間的コストを排除するために、スピーチ部分データ構造体からスピーチ部分記録ヘッダがコピーされる。スピーチ部分データ構造体971は、ワード「people」を表す語彙記録970への両方向性リンクを含む。別のデータ構造体972は、ワード「people」に対して考えられるスピーチ部分の動詞を含むと共に、語彙記録970への両方向性リンクも含む。スピーチ部分データ構造体971及び972と語彙記録データ構造体970との間のリンクは、ルールがそれらの処理を、特定のワードに対して考えられる全てのスピーチ部分のセットに基づいて行えるようにする。ワード「people」に対する名詞のスピーチ部分記録914の場合に、この記録に適用されるルールは、ワード「people」が動詞も表し得ることを考慮する。

【0019】図10は、本発明のこの特徴によるチャートの更に別の変形を示している。図10は、考えられる意図されたワードに対するスピーチ部分データ構造体と、その潜在的に混同するワードに対する語彙記録データ構造体とのリンクを示す。図10は、名詞形態のワード「ad」に対するスピーチ部分データ構造体1052を「ad」に対する語彙記録1050に加えるところを示している。両スピーチ部分データ構造体1051及び1052は、ワード「ad」の語彙データ構造体1050に両方向性リンクされるので、「ad」に対する動詞のスピーチ部分記録1012に適用されるルールは、このワードに対して考えられる名詞のスピーチ部分とみなすことができる。上記のように、本発明によれば、考えられる意図されたワードの考えられるスピーチ部分に対するスピーチ部分データ構造体は、ルールにより結合されるべきワードに対して考えられる他のスピーチ部分に基づいて作用するルールの適用を容易にするために、各々の考えられる混同するワードに対する語彙データ構造体にリンクされるのが好ましい。図11は、本発明のこの特徴により、名詞のスピーチ部分記録1116がステップ806においてチャートに追加されたときに、考えられる意図されたワード「ad」に対する潜在的なスピーチ部分の名詞を含むスピーチ部分データ構造

体1152へのポインタを含むことを示している。

【0020】完全なパズルの発生に貢献しようにないルールの適用を防止することによってパージングの効率を高めるのに加えて、考えられる意図されたワードに対してリンクされるスピーチ部分記録を参照するルールを使用することは、潜在的に混同するワードが考えられる意図されたワードと実際に混合する場合にパージが潜在的に混同するワードを用いて入力テキストの見掛け上正しい完全なパズルを発生するのを防止することができる。

これは、混同するワード「form」が意図されたワード「from」に代わって使用される次の例示的なセンテンスについて言えることである。Angela departed from Seattle. 考えられる意図されたワードの考えられるスピーチ部分を考慮しないルールを用いると、あるパージは、動詞句が動詞句

「departed」と名詞句「form Seattle」から形成されるこのセンテンスの完了パズルを発生する。この完了パズルは、「form Seattle」が動詞「departed」の有効な目的語でないから、実際には正しくない。しかしながら、考えられる意図されたワードの考えられるスピーチ部分を考慮するルールを用いると、パージは、この誤った完了パズルを回避することができる。この場合に、動詞句と動詞句の目的語である名詞句を結合して別の動詞句にするルールを適用することは、名詞句の「前修飾語」（即ち、名詞句のメインワード「Seattle」の前に生じるワード「form」）又はその他の考えられる意図されたワードがスピーチの前置詞部分を表し得るときに、スピーチの前置詞部分が動詞の目的語の前に意図される可能性が大きい場合は、阻止される。考えられる意図されたワード「from」は、スピーチの前置詞部分をもつことができ、そして前置詞のスピーチ部分記録は、潜在的に混同するワード「form」のスピーチ部分記録にリンクされるので、このルールの適用は阻止され、パージが上記の誤った完了パズルを形成しないよう防止し、これにより、パージの出力の精度を改善する。

【0021】図12及び13は、潜在的に混同するワードのスピーチ部分記録がチャートに追加されてそして暗示されたルールがステップ805によりそれらに適用された後に、別のワードに対するスピーチ部分記録をチャートに追加させる2つの好ましい実施形態を示している。

図12は、2つの異なるリスト即ち「待ち行列」である一次リスト1280及び二次リスト1290からスピーチ部分記録がチャート120に追加される実施形態を示す。入力テキストに含まれたワードに対するスピーチ部分記録は、一次リスト1280に記憶される。この一次リストは、入力テキストに現れるワード「the」、

「add」、「convinced」及び「people」に対するスピーチ部分記録を含む。二次リスト1290は、考えられる意図されたワードに対するスピーチ

部分記録を含む。二次リスト1290は、別のワード「ad」に対する名詞のスピーチ部分記録を含むことが明らかである。この実施形態では、ワードは、先ず、一次リストからチャートに追加される。暗示されたルールが適用された後に、パージは、二次リストからスピーチ部分記録をチャートに追加する。好ましい実施形態において、スピーチ部分記録を二次リストからチャートに追加することは、先ず、二次リストから一次リストへそれらを移動し、次いで、それらを一次リストからチャートへ追加する一方、一次リストに現れる新たに暗示されるルールを適用することを含む。この解決策は、入力テキストのワードと一般に混同するところの考えられる意図されたワードがチャートに追加される前に、入力テキストに含まれたワードに対するスピーチ部分記録からパズルを構成できるようにする。

【0022】図13は、スピーチ部分記録が単一のリスト1370からチャート1300に追加される別の実施形態を示す。リスト1370は確率リストであり、完了パズルのリーフを最終的に構成する各スピーチ部分記録の確率に基づいて分類される。確率指向パージの詳細な説明については、参考としてここに取り上げる「統計学的手法をルールに基づく自然言語パージへとブートストラップするための方法及びシステム(METHOD AND SYSTEM FOR BOOTSTRAPPING STATISTICAL PROCESSING INTO A RULE-BASED NATURAL LANGUAGE PARSER)」と題する米国特許出願第08/265,845号を参照されたい。スピーチ部分記録は、成功確率パズルのリーフを構成する確率の下降順に確率リストからチャートに追加される。これらの確率は、「適用優先値」とも称されるが、入力テキストセグメントの代表的集成に対して完成したパズルにおける各スピーチ部分記録の出現を統計学的に分析することによって発生されるのが好ましい。例えば、スピーチ部分記録1374及び1375に関連して示された統計データは、ワード「people」を含む入力テキストセグメントにおいて、ワード「people」が入力セグメントの完全なパズルにおいて名詞を表すときが78%であり、一方、動詞を表すケースが13%であることを示している。この実施形態において、考えられる意図されたワードに対するスピーチ部分記録には、比較的小さな確率が指定される。それらを処理の終わり付近でチャートに追加させるのが好ましい。これは、多数の方法で行われる。ワード「ad」は入力テキストに実際に生じないので、入力テキストに「ad」が現れるときにワード「ad」の名詞形態が完全なパズルのリーフを形成する確率は、低くすることができる。或いは又、ワード「ad」ではなくワード「add」が入力テキストに現れるときにワード「ad」の名詞形態が完成パズルのリーフを形成する確率を計算するように個別の統計データを維持することもできる。いずれにせよ、代替ワード「ad」に対す



るスピーチ部分記録は、潜在的に混同するワード「add」に対するスピーチ部分記録の後にチャートに追加される。最後に、考えられる意図されたワードに対するスピーチ部分記録の確率は、確率リストの最少確率以下にセットすることができる。

【0023】更に別の好ましい実施形態では、潜在的に混同するワードに対するスピーチ部分記録がチャートに追加された後に代替えワードに対するスピーチ部分記録をチャートに追加させる2つの解決策を組み合わせて、潜在的に意図されたワードに対するスピーチ部分記録を二次リストに記憶し、そして全てのルールと、一次リストで終わりとなるスピーチ部分記録とをそれらの確率により順序付けする。

【0024】ユーザは、潜在的に混同するワード、又はその潜在的に混同するワードに置き換わる考えられる意図されたワードのリストを変更するようにパーザを構成できるのが好ましい。ユーザは、図4に示された混同し得るワードのテーブルを変更することによりこれを行う。しかしながら、あるユーザは、混同し得るワードのリストの簡単な表示を変更できることを望む。図14は、簡単化された混同し得るワードのファイル1400を示す。行1401ないし1406は、潜在的に混同するワードの1つのセットに各々対応する。特に指示のない限り、1つの行に一緒に現れるワードは、全て、互いに混同する。例えば、行1401は、ワード「add」がワード「add」と混同し得るその逆もあり得ることを示す。又、この混同し得るワードのファイルは、潜在的に混同するとみなしてはならないワードの前にハイフオン（-）記号を置くことによりユーザが一方の混同関係を指定できるのが好ましい。例えば、行1401におけるワード「can't」の前のハイフオン記号は、ワード「cant」がワード「can't」と混同することはあるが、ワード「can't」がワード「cant」と潜在的に混同されないことを示す。又、この混同し得るワードのファイルは、潜在的に混同するセットのワードは、それらが同じスピーチ部分、数及び時制を有しているも、互いに置き換えられることをユーザがアスタリスク（\*）記号で指定できるのが好ましい。例えば、混同し得るセット1402の前のアスタリスク記号は、「can't」及び「cant」の両方のワードがスピーチの動詞部分を有しているも、ワード「can't」をワード「cant」に置き換えられることを示す。パーザは、ユーザがパーザの動作を構成するように混同し得るワードのファイル1400を変更できると共に、混同し得るワードのファイルを、図4に示す混同し得るワードのテーブルのようなパーザにより容易に適用される形態へと変換できるのが好ましい。

【0025】好ましい実施形態を参照して本発明を説明したが、当業者であれば、本発明の範囲から逸脱せず、種々の変更や修正がなされ得ることが明らかである

う。例えば、上記以外の機構を使用して、一般に混同するワードに対するスピーチ部分記録をチャートに導入することができる。更に、上記した本発明の実施形態は、コンピュータプログラミング言語やテキスト形成言語のような人為的言語のテキストをパースするのに容易に適用できる。

#### 【図面の簡単な説明】

【図1】本発明が好ましく動作する汎用コンピュータシステムの高レベルブロック図である。

【図2】入力テキストに生じるワードに対してパーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。

【図3】チャートの内容により暗示されるルールをパーザが適用するところを示すチャート図である。

【図4】入力テキストにおける潜在的に混同するワードを識別するのに使用される例示的な混同ワードテーブルを示すテーブル図である。

【図5】入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対し、パーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。

【図6】考えられる意図されたワードに対するスピーチ記録の部分がチャートに追加された後に、チャートの内容により暗示されたルールをパーザが適用するところを示すチャート図である。

【図7】パーザを用いた文法チェッカーの視覚的ユーザインターフェイスを示すスクリーン図である。

【図8】一般に混同するワードを識別しそして分析しながら入力テキストをパースするために適用パーザにより好ましく実行される高レベルステップを示すフローチャートである。

【図9】辞書からの語彙記録を含む図2の別の図であって、ステップ801の実行後且つルールを適用する前のチャートの内容を示す図である。

【図10】代替えワードに対するスピーチ部分データ構造体を、その潜在的に混同するワードに対する語彙記録データ構造体にリンクするところを示す図である。

【図11】考えられる意図されたワードに対するスピーチ部分記録を追加した後のチャートの内容を示す図である。

【図12】スピーチ部分記録が2つの異なるリストからチャートに追加される実施形態を示す図である。

【図13】スピーチ部分記録が、確率でランク付けされた単一リストからチャートに追加される別の実施形態を示す図である。

【図14】混同し得るワードの簡単化されたファイルを示す図である。

#### 【符号の説明】

100 コンピュータシステム

110 中央処理ユニット

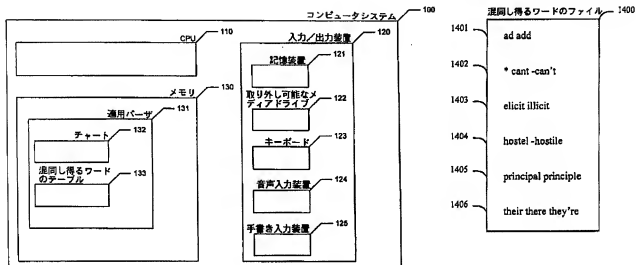
- 120 入力/出力装置  
 121 記憶装置  
 122 取り外し可能なメディアドライブ  
 123 キーボード  
 124 音声入力装置  
 125 手書き入力装置  
 130 コンピュータメモリ

- \* 131 パーザ  
 132 チャート  
 133 混同し得るワードのテーブル  
 200 パーザのチャート  
 201 入力ストリング  
 211-215 スピーチ部分記録

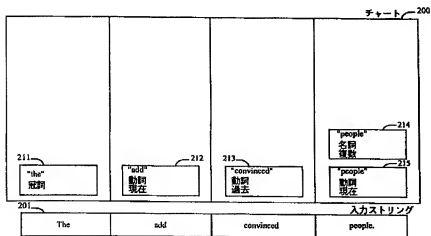
\*

【図1】

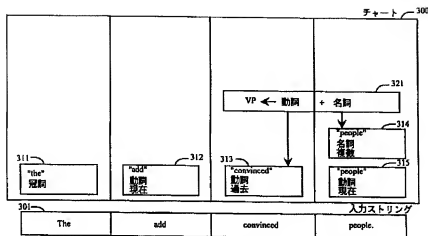
【図14】



【図2】



【図3】

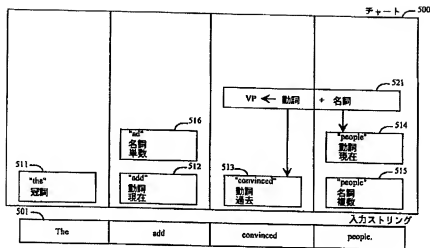


【図4】

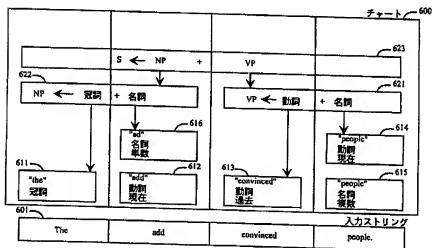
混同し得るワードのテーブル 400

| 潜在的に混同するワード | 考えられる意図されたワード      |
|-------------|--------------------|
| ad          | add 401            |
| add         | ad 402             |
| cant        | can't 403          |
| elicit      | illicit 404        |
| hostel      | hostile 405        |
| illicit     | elicit 406         |
| principal   | principle 407      |
| principle   | principal 408      |
| their       | there, they're 409 |
| there       | their, they're 410 |
| they're     | their, there 411   |

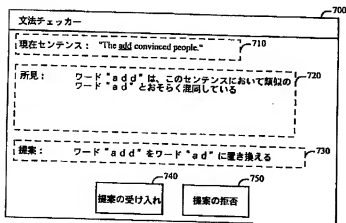
【図5】



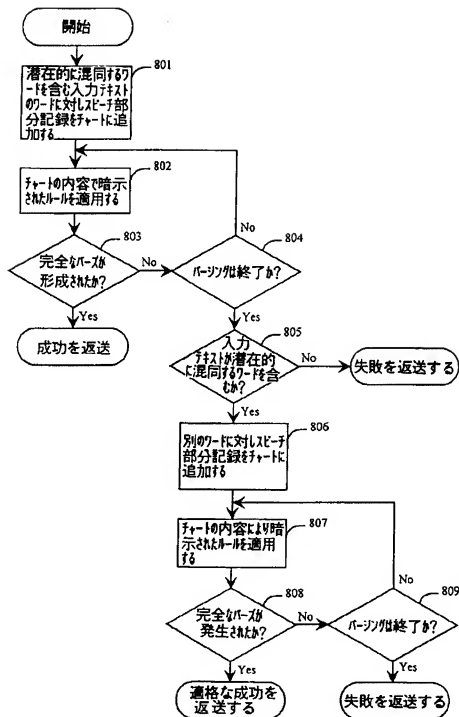
【図6】



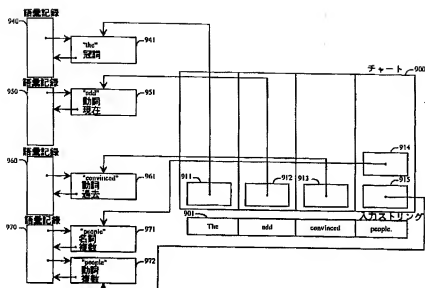
【図7】



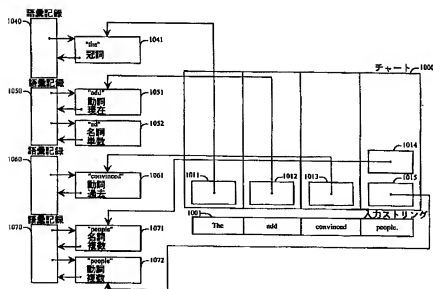
【図8】



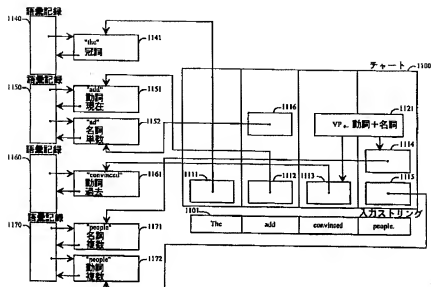
【図9】



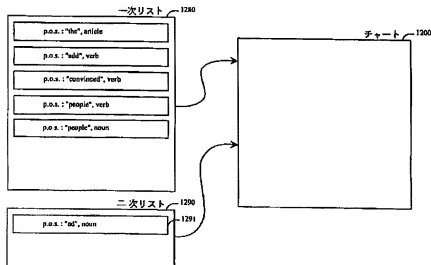
【図10】



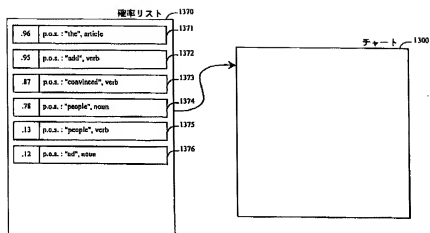
【図11】



【図12】



【図13】



フロントページの続き

(72)発明者 ジョージ イー ヘイドーン  
 アメリカ合衆国 ワシントン州 98008  
 ベルビュー ワンハンドレッドアンドシッ  
 クスティフィフス プレイス ノースイー  
 スト 3211